

Topic Browsing for Research Papers with Hierarchical Latent Tree Analysis

Leonard K.M. Poon

The Education University of Hong Kong
10 Lo Ping Road, Hong Kong, China
kmpoon@eduhk.hk

Nevin L. Zhang

The Hong Kong University of Science and Technology
Clear Water Bay Road, Hong Kong, China
lzhang@cse.ust.hk

Abstract

Academic researchers often need to face with a large collection of research papers in the literature. This problem may be even worse for postgraduate students who are new to a field and may not know where to start. To address this problem, we have developed an online catalog of research papers where the papers have been automatically categorized by a topic model. The catalog contains 7719 papers from the proceedings of two artificial intelligence conferences from 2000 to 2015. Rather than the commonly used Latent Dirichlet Allocation, we use a recently proposed method called hierarchical latent tree analysis for topic modeling. The resulting topic model contains a hierarchy of topics so that users can browse the topics from the top level to the bottom level. The topic model contains a manageable number of general topics at the top level and allows thousands of fine-grained topics at the bottom level. It also can detect topics that have emerged recently.

Introduction

Academic researchers often need to face with a large collection of research papers in the literature. This problem may be even worse for postgraduate students who are new to a field and may not know where to start. Researchers usually use keywords to search for related papers using a search engine. After reading some papers, they then try to group related papers together to discover the main topics in the field. This process can be time-consuming.

The approach above can be regarded as bottom-up approach. A top-down approach would be to start with topic hierarchy. Researchers can then pick a general topic and drill down to more specific topics. Papers related to any of the topics can be presented to the researchers when requested.

To allow the top-down approach, traditionally a taxonomy has to be defined manually. Papers can then be manually categorized according to the taxonomy. One problem with the traditional method is that it requires much effort. Besides, the topics in the taxonomy may not be able to keep up with recent development.

Topic models can be used to automate this process. They can be used to detect topics from a collection of documents

and categorize documents according to the detected topics. We use a recently proposed method called *hierarchical latent tree analysis* (HLTA) (Liu, Zhang, and Chen 2014; Chen et al. 2016) for topic modeling. Unlike Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), HLTA yields a hierarchy of topics. hLDA (Blei, Griffiths, and Jordan 2010) and nHDP (Paisley et al. 2015) are two extensions of LDA for producing topic hierarchy. However, HLTA has been recently shown to produce better quality of topics and topic hierarchy than the two LDA extensions (Chen et al. 2016).

To facilitate the top-down approach, we have developed an online catalog of research papers where the papers have been automatically categorized by a topic model built with HLTA. The catalog contains 7719 papers from the proceedings of two artificial intelligence conferences from 2000 to 2015. The resulting topic model contains a hierarchy of topics so that users can browse the topics from the top level to the bottom level. The topic model contains a manageable number of general topics at the top level and allows thousands of fine-grained topics at the bottom level. It also can detect topics that have emerged recently.

This paper is organized as follows. In the next section we review background of our work. We then explain HLTA using an example. Next, we describe the procedure for building the online catalog. In the following section, we show the results and observations obtained from the online catalog. After that, we conclude the paper.

Background

In topic modeling, documents are usually represented as *bags of words*. Consider a collection $\mathcal{D} = \{d_1, \dots, d_N\}$ of N documents. Suppose M words are included in the vocabulary $\mathcal{V} = \{w_1, \dots, w_M\}$. Each document d can be represented as a vector $d = (c_1, \dots, c_M)$, where c_i represents the count of word w_i occurring in the document d . The aim of topic modeling is to detect a number K of topics z_1, \dots, z_K among the documents \mathcal{D} . The number K can be given or learned. The topic model defines a distribution over words for each topic. A topic is often characterized by representative words based on the distribution.

Latent Dirichlet Allocation (LDA) is a popular method for topic modeling (Blei, Ng, and Jordan 2003; Blei 2012). LDA assumes each document d to belong to the K topics according to a distribution $P(Z = z_k | d)$ over the topics.

In other words, $\sum_1^K P(z_k|d) = 1$. This kind of model is known as *mixed membership model*. For each topic z_k , LDA defines a conditional distribution $P(w_i|z_k)$ over words w_i . A topic z_k can then be characterized by the most probable words according to $P(w_i|z_k)$.

Latent Tree Models

A *latent tree model* (LTM) is a tree-structured probabilistic graphical model (Zhang 2004; Chen et al. 2012). Figure 2a shows an example of LTM. When an LTM is used for topic modeling, the leaf nodes represent the observed word variables \mathbf{W} , whereas the internal nodes represent the unobserved topic variables \mathbf{Z} . All variables are binary. Each word variable $W_i \in \mathbf{W}$ indicates the presence or absence of the word $w_i \in \mathcal{V}$ in a document. Each topic variable $Z_i \in \mathbf{Z}$ indicates whether a document belongs to the i -th topic.

For technical convenience, we often root an LTM at one of its latent nodes and regard it as a Bayesian network (Pearl 1988). Then all the edges are directed away from the root. The numerical information of the model includes a marginal distribution for the root and one conditional distribution for each edge. For example, edge $Z1314 \rightarrow \text{dean}$ is associated with probability $P(\text{dean}|Z1314)$. The conditional distribution associated with each edge characterizes the probabilistic dependence between the two nodes that the edge connects. The product of all those distributions defines a joint distribution over all the latent variables \mathbf{Z} and observed variables \mathbf{W} . Denote the parent of a variable X as $pa(X)$ and let $pa(X)$ be an empty set when X is the root. Then the LTM defines a joint distribution over all observed and latent variables as follows:

$$P(\mathbf{W}, \mathbf{Z}) = \prod_{X \in \mathbf{W} \cup \mathbf{Z}} P(X|pa(X))$$

Given a document d , the values of word variables \mathbf{W} are observed. Use $d = (w_1, \dots, w_M)$ to denote also those observed values. Whether a document d belongs to a topic $Z \in \mathbf{Z}$ can be determined by the probability $P(Z|d)$. The LTM gives a *multi-membership model* since a document can belong to multiple topics. Unlike in LDA, the topic probabilities $P(Z|d)$ in LTM do not necessarily sum to one.

Hierarchical Latent Tree Analysis

For topic modeling, an LTM has to be learned from the document data \mathcal{D} . This requires learning the number of topic variables, the connection between the variables, and the probabilities in the model.

We use the method PEM-HLTA proposed by Chen et al. (2016) to build LTMs for topic modeling. The method builds LTMs level by level and is thus also known as *hierarchical latent tree analysis* (HLTA). In this section, we use an example to illustrate the main ideas of PEM-HLTA. Readers are referred to the original paper for details.

As an example, consider a data set that contains the 24 word variables in Figure 2a. PEM-HLTA is an iterative procedure that builds one level of model in each iteration. In the first iteration, it partitions the 24 word variables in 6 clusters (Figure 1c). The clusters are *unidimensional* in the sense that

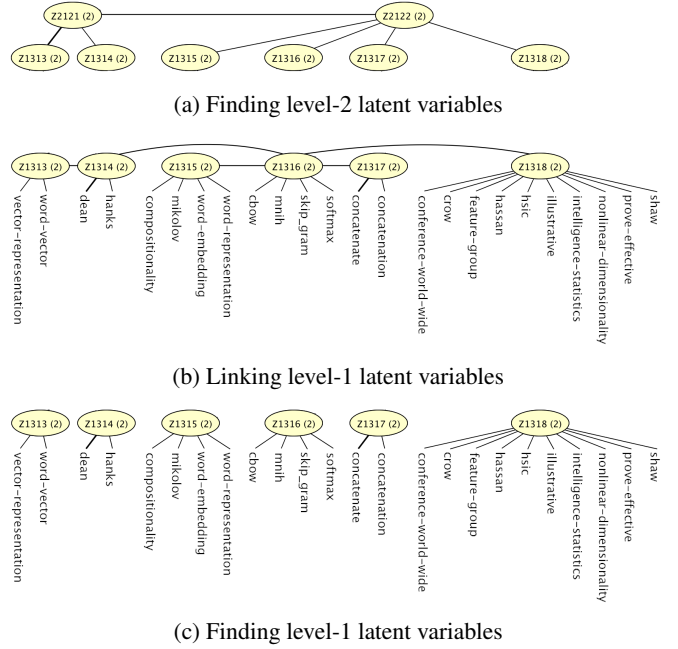


Figure 1: Illustrative example of PEM-HLTA.

the co-occurrences of words in each cluster can be properly modeled using a single latent variable. A latent variable is introduced for each cluster to form model in which all variables in a cluster are connected to the latent variable. We metaphorically refer to those models corresponding to the clusters as islands and the latent variables in them as level-1 latent variables.

The next step is to link up the 6 islands. This is done by estimating the mutual information (MI) (Cover and Thomas 2006) between every pair of latent variables and building a Chow-Liu tree (Chow and Liu 1968) over them, so as to form an overall model (Liu et al. 2015). The result is the model in the Figure 1b.

To build the next level of model, inference is carried out to compute the posterior distribution of each level-1 latent variable for each document. The document is assigned to the state with the maximum posterior probability, resulting in a data set over the level-1 latent variables. The data set is used as the input for the next iteration. In the second iteration, the level-1 latent variables are partitioned into 2 groups. The 2 islands are linked up to form the model shown in Figure 1a. The model in Figure 1a is then stacked on the model in Figure 1b and a new data set over level-2 latent variables are computed. The iteration continues until the model in Figure 2a is obtained.

Intuitively, the co-occurrence of words are captured by the level-1 latent variables, whose co-occurring patterns are captured by higher level latent variables. Then a topic hierarchy can be extracted, with topics on top more general and topics at the bottom more specific.

Algorithm 1 Find-NGrams(\mathcal{D}, m, M)

Input: \mathcal{D} – Document collection, m – maximum value of n , M – number of tokens to be selected.

Output: A document collection with individual words replaced by selected n-grams.

- 1: Find individual words for each $d \in \mathcal{D}$.
 - 2: Set \mathcal{V} to be the set of the M selected words.
 - 3: **for** $n = 2$ to m **do**
 - 4: For each $d \in \mathcal{D}$, form an n -gram for each pair of consecutive tokens t_1, t_2 in d if $t_1, t_2 \in \mathcal{V}$ and if the resulting n -gram has a length of n .
 - 5: Denote the set of newly formed n -grams by \mathcal{U} .
 - 6: Set \mathcal{V} to be the set of M tokens $t \in \mathcal{V} \cup \mathcal{U}$ that have the highest $\text{tf-idf}(t)$.
 - 7: For each $d \in \mathcal{D}$, replace all pairs of consecutive tokens that can be used to form an n -gram in \mathcal{V} . Let \mathcal{D} be the collection of documents after replacement.
 - 8: **end for**
 - 9: **return** \mathcal{D} .
-

Building Online Catalog with HLTA

In this section, we describe the procedure for building an online catalog of documents with HLTA. The procedure starts with each document contained in a PDF file.

Extract Text

Given a PDF file, we extract the text content using Apache PDFBox.¹ We remove hyphenation from the extracted text. After that we use Stanford Core NLP (Manning et al. 2014) for sentence splitting and lemmatization.²

The normalize the words, we convert all letters to lowercase. We also remove accents and ligatures using the `java.text.Normalizer` class in the Java library. We use underscore to replace all non-alphanumeric characters and starting digits in a word. We remove stop words³ and words with fewer than 4 characters.

Convert Data

After text is extracted, we compute the term frequency and document frequency. The term frequency $\text{tf}(w, d)$ is defined as the number of occurrences of a word w in document d . The document frequency $\text{df}(w)$ is defined as the number of documents that contain the word w .

We remove words that occur in more than 25% of documents. In other words, a word w is removed if $\text{df}(w) \geq 0.25N$, where N is the number of documents. Given a number M , we select the M words with highest TF-IDF, which is given by:

$$\text{tf-idf}(w) = \frac{1}{\ln \text{df}(w)} \sum_{d \in \mathcal{D}} \text{tf}(w, d).$$

¹<http://pdfbox.apache.org/>

²<http://stanfordnlp.github.io/CoreNLP/>

³<http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop>

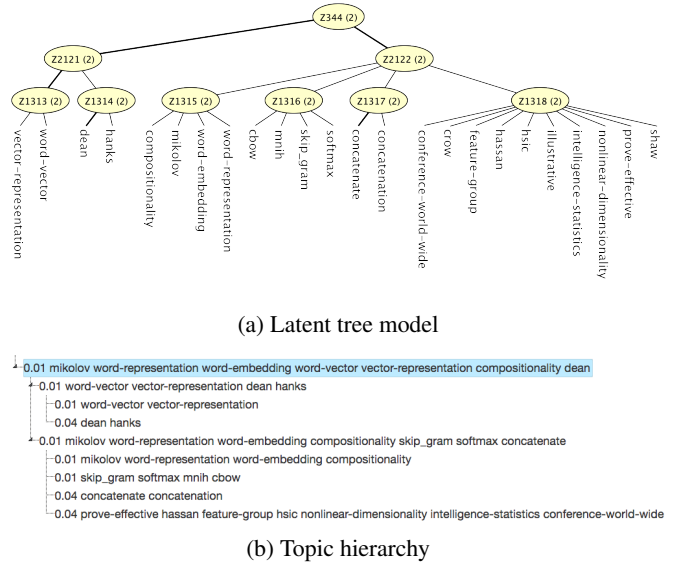


Figure 2: Illustration of topic extraction. Each shaded node in (a) corresponds to a row in (b). Both of them represent a topic. The topic hierarchy (b) is extracted from the model (a).

After selecting the words, we can represent each document d as a vector $d = (w_1, \dots, w_M)$, where w_i corresponds to one of the selected words and its value indicates the presence ($w_i = 1$) or absence ($w_i = 0$) of the word in d .

Inspired by Deng et al. (2016), we consider also n-grams in addition to individual words. For a given value of m , we consider 1-grams, 2-grams, and so on up to m -grams. We also use the term *tokens* to refer to the n-grams. The leaf nodes of the model in Figure 2a show some examples of n-grams. The examples of 2-grams include `vector-representation` and `word-vector`, and an example of 3-gram is `conference-world-wide`.

The n-grams in a collection of documents \mathcal{D} can be found by Algorithm 1. After running the algorithm, each document can be converted to a binary vector similarly as above.

The tree structure of LTM limits a word variable to be connected to exactly one parent. The inclusion of n-grams mitigates this limitation. For example, the word `bayesian` may be related to `statistics` and the word `network` to `social networks` or `communities`. The 2-gram `bayesian-network` means a class of graphical models and has a different meaning from the individual words. By including 2-grams, the three tokens are allowed to three different parent nodes that are more closely to their respective meanings.

Building Model

After converting the documents to binary vectors, we run PEM-HLTA using those vectors as input data. Note that PEM-HLTA automatically determines the number of levels of latent variables and the number of latent variables in each level. The result of PEM-HLTA is an LTM.

Extract Topic Hierarchy

An LTM defines a tree structure of nodes. Each internal node represents a topic. Since the model defines a joint distribution over the variables, we can compute the MI between every pair of variables. For each topic, we compute the MI between each descendent word variable and the topic variable. We then pick the descendent words with highest MI to characterize the topic.

As an example, consider extracting topics from the model in Figure 2a. Each shaded node represents a topic. Suppose we want to characterize Z344. The descendent word variables are `vector-representation`, `word-vector`, ..., `shaw`. We compute the MI between Z344 and each of those word variables. The 7 words with highest MI are shown in the shaded row in Figure 2b. The row corresponds to the topic represented by Z344. The second and third rows correspond to the topics represented by Z2121 and Z1313 respectively. The topic extraction from the model in Figure 2a results in the topic hierarchy shown in Figure 2b.

In addition to finding the characterizing words, we also estimate the size of each topic Z , which is given by the marginal probability $P(Z)$. It shows how often a topic is estimated to occur in a document collection. In Figure 2b, the number on each row indicates the topic size.

Build Online Catalog

The LTM can be used to classify the documents according to the topics detected. A document d is assigned to topic Z if $P(Z = 1|d) > 0.5$. We use a webpage to display this information. On the webpage, a topic hierarchy similar to Figure 2b is shown. The hierarchy is built with a jquery plugin called jsTree.⁴ When a topic is clicked, a list of documents belonging to that topic will be shown.

Results and Observations

We have built an online catalog of research papers using the method discussed above. The papers were obtained from the proceedings of two AI conferences, namely AAAI Conference on Artificial Intelligence and International Joint Conference on Artificial Intelligence. Proceedings between year 2000 and 2015 were used. The resulting collection contains 7719 papers. We considered n -grams for $n = 1, 2, 3$ and selected 10,000 tokens based on TF-IDF.

Online Catalog and Source Code

The online catalog can be accessed from the URL <http://goo.gl/gtDJC8>. A screenshot is shown in Figure 3. The program for building the online catalog was written in Scala and Java. The source code can be obtained from <https://github.com/kmpoon/hlta>.

Topic Hierarchy

The topic hierarchy extracted from the LTM obtained contains 7 levels of topics. Table 1 lists the number of topics for each level.

⁴<https://www.jstree.com>

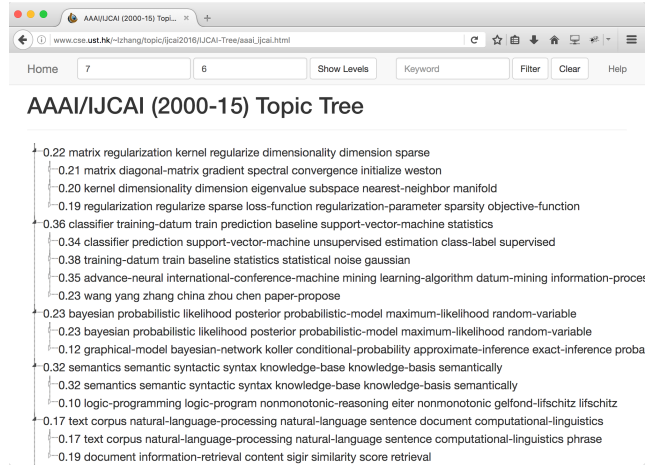


Figure 3: Screenshot of online catalog.

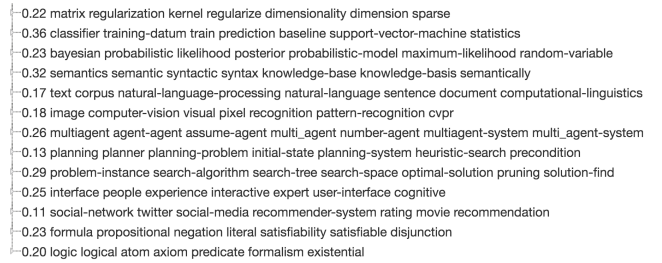


Figure 4: Top level topics

Level	No. of Topics
7	13
6	32
5	75
4	179
3	436
2	1173
1	3084
total	4992

Table 1: Number of topics detected for each level.

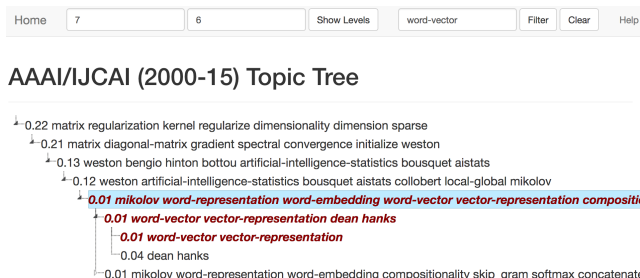


Figure 5: Searching topics with keyword word-vector.

The hierarchy contains 13 top-level topics (Figure 4). The 1st topic is about kernel methods, matrix factorization, dimensionality reduction. The 2nd one is about classifier and support vector machines. The 3rd one is about probabilistic models and Bayesian methods. The 4th to 7th topics are about knowledge base, natural language processing, computer vision, and agent systems, respectively. The 8th one is about planning and heuristic search. The 9th one is search. The 10th one is about to people and user interface. The 11th one is about social networks and recommender systems. The last two are about satisfiability and logic respectively.

Features

The online catalog provides some useful features for topic browsing.

Expanding and Collapsing Topics. The topic hierarchy webpage allows expanding a topic node to see its child topics or collapsing a topic node to hide its child topics. This allows users to navigate among more general topics and more specific topics. For example, if we expand the upper 5 topics in Figure 4, we see some more specific topics as in Figure 3. The bottom part of Figure 3 shows that the top-level topic about natural language processing can be divided into two more specific topics, one about natural language processing and one about information retrieval.

Levels of Topics. Users can specify a range of levels of topics to show at the top of the webpage (Figure 3). By default the top two levels of topics are shown.

Keyword Search. Users can enter a keyword at the top of the webpage to search for topics containing that keyword. Topics with the keyword are highlighted (Figure 5).

Document List. A list of document belonging to a topic can be shown when a topic node is clicked (Figure 6). The documents are sorted in descending order of membership as indicated by $P(Z|d)$. On this page, the numbers of documents belonging to the topic for each year are also shown in a table. We see that this topic about word-representation is emerging recently. It has only 9 documents from 2000 to 2013 but has 9 documents in 2014 and 58 documents in 2015.

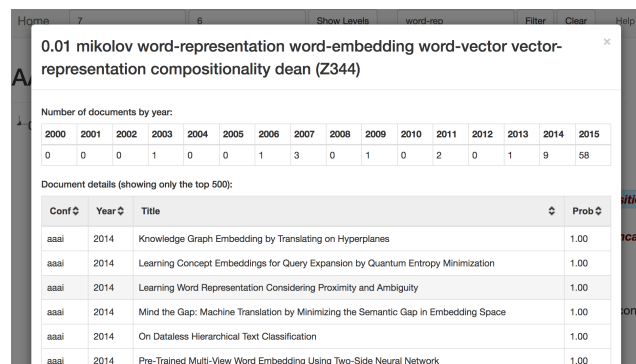


Figure 6: List of papers belonging to topic Z344.

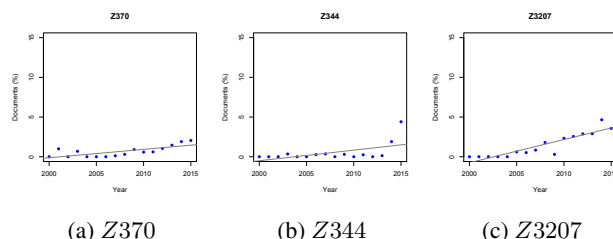


Figure 7: Proportion of documents belonging to a topic for each year.

Trends

For each document d , we can find the year of d and compute the topic indicator based on $P(Z|d)$. Hence, we can build a linear regression model using topic indicator as a predictor variable the year variable as response variable. We can then use the regression coefficient to estimate the trend of each topic. The trend allows researcher new to a field to consider whether a topic is worth for new work.

Table 2 shows the top three level-3 topics with an upward trend. The trends are supported by the increasing proportion of documents of the topics as shown in Figure 7.

Related Works

Topic browsing tools have been built based on topic models. Many tools use LDA for topic modeling (Gardner et al. 2010; Chaney and Blei 2012; Snyder et al. 2013; Sievert and Shirley 2014). They do not show any hierarchy of topics. Smith, Hawes, and Myers (2014) attempts to build a tool with a topic hierarchy by recursively splitting and re-modeling a corpus based on LDA. Unlike HLTA, the topic model does not have a strong statistical basis.

Conclusions

We present an online tool that allows users to browse topics from a hierarchy. The topic hierarchy is built using the recently proposed PEM-HLTA. The tool allows users to show documents related to each topic. The tool provides several features for easy browsing.

In the future, we will consider using more sophisticated way to detect n-grams. We may also include the hyperlinks

Z370	hash-function hash indyk hashing hash-method binary-code hash-code
Z344	mikolov word-representation word-embedding word-vector vector-representation compositionality dean
Z3207	marketing spread viral diffusion kleinberg-tardos kempe influence-maximization

Table 2: Top three level-3 topics with an upward trend.

to the papers in the document list for each topic. The online catalog currently takes some time to load in a web browser. We will consider storing the data in a database so that the loading time can be substantially reduced.

Acknowledgment

We thank Peixian Chen for the implementation of PEM-HLTA and Zichao Li for downloading the PDF files of the research papers. The work in this paper was supported by the Education University of Hong Kong under project RG90/2014-2015R and Hong Kong Research Grants Council under grant 16202515.

References

- [2010] Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested Chinese restaurant process and Bayesian non-parametric inference of topic hierarchies. *Journal of the ACM* 57(2):7:1–7:30.
- [2003] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.
- [2012] Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.
- [2012] Chaney, A. J.-B., and Blei, D. M. 2012. Visualizing topic models. In *International AAAI Conference on Web and Social Media*.
- [2012] Chen, T.; Zhang, N. L.; Liu, T.; Poon, K. M.; and Wang, Y. 2012. Model-based multidimensional clustering of categorical data. *Artificial Intelligence* 176:2246–2269.
- [2016] Chen, P.; Zhang, N. L.; Poon, L. K. M.; and Chen, Z. 2016. Progressive EM for latent tree models and hierarchical topic detection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.
- [1968] Chow, C. K., and Liu, C. N. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* 14(3):462–467.
- [2006] Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, 2nd edition.
- [2016] Deng, K.; Bol, P. K.; Li, K. J.; and Liu, J. S. 2016. On the unsupervised analysis of domain-specific chinese texts. *Proceedings of the National Academy of Sciences of the United States of America* 113(22):6154–6159.
- [2010] Gardner, M. J.; Lutes, J.; Lund, J.; Hansen, J.; Walker, D.; Ringger, E.; and Seppi, K. 2010. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*.
- [2015] Liu, T.-F.; Zhang, N. L.; Chen, P.; Liu, A. H.; Poon, L. K.; and Wang, Y. 2015. Greedy learning of latent tree models for multidimensional clustering. *Machine Learning* 98(1–2):301–330.
- [2014] Liu, T.; Zhang, N. L.; and Chen, P. 2014. Hierarchical latent tree analysis for topic detection. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)*, volume 8725 of *Lecture Notes in Computer Science*. Springer. 256–272.
- [2014] Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60.
- [2015] Paisley, J.; Wang, C.; Blei, D. M.; and Jordan, M. I. 2015. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(2):256–270.
- [1988] Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufmann Publishers.
- [2014] Sievert, C., and Shirley, K. E. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.
- [2014] Smith, A.; Hawes, T.; and Myers, M. 2014. Hierarchy: Interactive visualization for hierarchical topic models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 71–78.
- [2013] Snyder, J.; Knowles, R.; Dredze, M.; Gormley, M. R.; and Wolfe, T. 2013. Topic models and metadata for visualizing text corpora. In *Proceedings of the NAACL HLT 2013 Demonstration Session*, 5–9.
- [2004] Zhang, N. L. 2004. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research* 5:697–723.